Artif Intell Rev (2014) 42:275–293
DOI 10.1007/s10462-012-9338-y

Mixture of experts: a literature survey

Saeed Masoudnia · Reza Ebrahimpour

Published online: 12 May 2012

Abstract Mixture of experts (ME) is one of the most popular and interesting combining methods, which has great potential to improve performance in machine learning. ME is established based on the divide-and-conquer principle in which the problem space is divided between a few neural network experts, supervised by a gating network. In earlier works on ME, different strategies were developed to divide the problem space between the experts. To survey and analyse these methods more clearly, we present a categorisation of the ME literature based on this difference. Various ME implementations were classified into two groups, according to the partitioning strategies used and both how and when the gating network is involved in the partitioning and combining procedures. In the first group, The conventional ME and the extensions of this method stochastically partition the problem space into a number of subspaces using a special employed error function, and experts become specialised in each subspace. In the second group, the problem space is explicitly partitioned by the clustering method before the experts' training process starts, and each expert is then assigned to one of these sub-spaces. Based on the implicit problem space partitioning using a tacit competitive process between the experts, we call the first group the mixture of implicitly localised experts (MILE), and the second group is called mixture of explicitly localised experts (MELE), as it uses pre-specified clusters. The properties of both groups are investigated in comparison with each other. Investigation of MILE versus MELE, discussing the advantages and disadvantages of each group, showed that the two approaches have complementary features. Moreover, the features of the ME method are compared with other popular combining methods, including boosting and negative correlation learning methods. As the investigated methods have complementary strengths and limitations, previous researches that attempted to combine their features in integrated approaches are reviewed and, moreover, some suggestions are proposed for future research directions.

S. Masoudnia
School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran

R. Ebrahimpour (✉)
Department of Electrical and Computer Engineering, Brain and Intelligent Systems Research Laboratory, Shahid Rajaee Teacher Training University, P.O. Box:16785-163, Tehran, Iran
e-mail: ebrahimpour@ipm.ir

Keywords Classifier combining · Mixture of experts · Mixture of implicitly localised experts · Mixture of explicitly localised expert

1 Introduction

Among the conventional classification methods, i.e., support vector machines, fuzzy systems and neural networks (NNs) methods are widely used in pattern recognition problems. In comparison with each other, these methods have several advantages and disadvantages in solving wide range of various classification problems (Kecman 2001). However, both empirical studies and specific machine learning applications verify that a given classification method outperforms all others for a particular problem or for a specific subset of the input data, but it is abnormal to find a single method achieving the best results on the overall problem domain (Dietterich 2000). As a consequence combining classifiers try to exploit the local different behavior of the base classifiers to improve the accuracy and the reliability of the overall classification system. There are also hopes that if a classifier fails, the overall system can recover the error (Kotsiantis 2011a).

As the other view-point, combining classifiers is an approach to improve the performance in classification (Kotsiantis et al. 2006; Lorena et al. 2008; Rokach 2010) particularly for complex problems such as those involving limited number of patterns, high-dimensional feature sets, and highly overlapped classes (Tran et al. 2011; Kotsiantis 2011b,a). Combining neural network (NN) methods have two major components, i.e., a method to create individual NN experts and a method for combining NN experts. Both theoretical and experimental studies (Tumer and Ghosh 1996) have shown that combining procedure is the most effective when the experts' estimates are negatively correlated; but this procedure is moderately effective when the experts are uncorrelated and only mildly effective when the experts are positively correlated. Therefore, more improved generalization ability can be obtained by combining the outputs of NN experts which are accurate and their errors are negatively correlated (Jacobs 1997; Hansen 2000).

There are a number of alternative approaches that were used to produce negatively correlated NNs for the ensemble. These approaches include varying the initial random weights of NNs, varying the topology of NNs, varying the algorithm employed to train NNs, and varying the training sets of NNs. It is argued that training NNs using different training sets is likely to produce more uncorrelated errors than other approaches (Sharkey and Sharkey 1997). The two popular algorithms to construct ensembles that train individual NNs independently and sequentially, using different training sets are bagging (Breiman 1996) and boosting (Schapire 1990) algorithms, respectively. Negative correlation learning (NCL) (Liu and Yao 1999a) and mixture of experts (ME) (Jacobs et al. 1991b), as the other combining methods, employ special error functions to train NNs simultaneously producing negatively correlated NNs.

It was shown that in contrast to common combining methods that produce unbiased experts whose estimation errors are uncorrelated, the ME method produce biased experts whose estimates are negatively correlated (Jacobs 1997). This combining method has special features compared to common combining methods. The ME method is established based on Divide-and-Conquer (D&C) principle (Jacobs et al. 1991b). In this method, the problem space is partitioned stochastically into a number of subspaces through special employed error function, experts become specialized on each subspace. This method uses a gating network to manage this process, which trains together with the experts. The gating network during the training of the experts, with respect to difference in the experts' efficiencies in

the different sub-spaces co-operate in the partitioning of problem, simultaneously. In this method, instead of assigning a set of fixed combinational weights to the experts, the gating network is used to compute these weights dynamically from the inputs, according to local efficiency of each expert. Based on these special features among various types of combining methods (Jacobs 1997; Polikar 2006), ME has attracted considerable attention in the literature of combining methods (Avnimelech and Intrator 1999; Ubeyli 2009; Xing and Hua 2008; Ebrahimpour et al. 2010).

In this paper, we present a review of ME literature. To analyse the basic features of ME, first the position of ME in various taxonomies of combining methods are described and the training algorithm of conventional ME are then investigated. Approaches using D&C in the ME method were implemented in different ways, and several methods were developed based on this model, which used different strategies to divide the problem between the experts (Jacobs et al. 1991a; Gutta et al. 2000; Tang et al. 2002; Ebrahimpour et al. 2011b). To survey and analyse the ME-based methods more clearly, we present a categorisation of the ME literature based on this difference. Various ME implementations were classified into two groups, according to the partitioning strategies used. The properties of both groups in ME literature are investigated in comparison with each other, discussing their advantages and limitations. To present a complete survey and moreover, suggest novel ideas for future research directions, the hybrid methods in which the strengths of ME are incorporated in the other combining methods are also reviewed.

The rest of this paper is organized as follows. Section 2, first presents the basic features of ME and its position in various taxonomies of combining methods. The training algorithm of conventional ME and its extension are then described. In the Sect. 3, first our proposed categorisation in the ME literature is presented. Next, Two groups of ME methods in the literature are investigated in comparison with each other, discussing their advantages and limitations. Section 4 first presents a review of common combining methods in terms of partitioning strategy to survey the hybrid methods proposed based on complementary features of ME. Section 5 concludes the paper and finally, in the Sect. 6, some suggestions are proposed for future research direction.

2 Background on ME

The ME method employed specific approaches as two components of its combining system. In order to create individual NN experts, this method uses special error function to localise the base experts in different distributions of data space. In this procedure, complex problem based on D&C approach is partitioned into a set of simpler sub-problems between the experts (Jacobs et al. 1991b). In the combiner component, ME employs an approach that can model the local competence of the experts in different distribution of data space according to each input data. To further explicate the ME literature, first the basic features of ME and its position in the taxonomy of combining methods are reviewed. Next, the conventional training algorithm of ME and mixture of MLP-experts, as one of the most applied implementations of ME, are also described.

2.1 Basic features of ME and its position in the taxonomy of combining methods

To describe the features of the ME method, the combining methods should be reviewed from various perspectives. We thus review and categorise the combining methods according to various criteria, and we also determine the ME position in the taxonomy of combining methods.

An important parameter in analysing the combining methods is how they combine the output from base experts. According to the existence of the training process in the combiner part of the ensemble, combining methods are classified into two categories. Several combiner methods require no training after classifiers in the ensemble are trained individually. Other combiners require additional training before, during or after training the individual classifiers, including the weighted average combiner, Stacked Generalisation (SG), AdaBoost and ME. The first group is called non-trainable, and the second is a trainable ensemble. The trainable combining methods, according to the influence of input data on the combining process, are classified into data-dependent and data-independent ensemble. In the first group, the input has explicit interference in the ensemble, as the combining weights are functions of the input, while the input does not influence the combining process in the second group. ME is considered the most famous method in the first class, and weighted averaging and SG are from the latter class (Kuncheva 2004).

From the other viewpoint, combining methods act based on two different strategies: fusion and selection. In classifier fusion, it is assumed that each ensemble member is trained on the whole feature space (Kittler et al. 1998), whereas in classifier selection, each member is assigned to learn a part of the feature space (Woods et al. 1997; Alpaydin and Jordan 1996). Therefore, in the former strategy, the final decision is made considering the decisions of all members, while in the latter strategy, the final decision is made by aggregating the decisions of one or a few of experts. There also exist combination schemes lying between the two pure strategies (Kuncheva 2002). Such a scheme; for example, is taking the average of outputs with coefficients which depend on the input x. Thus, the local competence of the experts, with respect to x, is measured by the weights. Then, more than one classifier is responsible for x and the outputs of all responsible classifiers are combined. The ME method is the most famous example of a scheme between selection and fusion (Kuncheva 2004).

According to these taxonomies of combining methods, ME is a data-dependent trainable combining method that, with respect to input, acts based either on the selection or fusion strategies.

2.2 Conventional ME

The ME method was introduced by Jacobs et al. (1991b). The authors examined the use of different error functions in the learning process for expert networks in the ME method. Jacobs et al. proposed making NNs into local experts for different distributions of data space; as a result, the increased diversity among the experts led to improvements in the performance of this method. Various error functions were then investigated with respect to a performance criterion.

In the first test, the following error function was used for the experts:

$$E = \left\| y - \sum_j g_j O_j \right\|^2 \tag{1}$$

where y and O_j are target vector and the output of expert j, respectively and g_j is the proportional contribution of expert j to the combined output vector.

According to an analysis of the derivation of this error function, the weights of each expert are updated based on the overall ensemble error rather than the errors of each expert. This strong coupling in the process of updating the weights of the experts engenders a high level of cooperation over the whole problem space and tends to employ almost all of the experts

for each data sample. This situation is inconsistent with the localisation of the experts in different data distribution.

The second error function analysed was the following:

$$E = \sum_j g_j \left\| y - O_j \right\|^2 \tag{2}$$

According to the derivation of this term, the weights of each expert are updated based on their own error in the prediction of the target and yields a complete output vector rather than a residual, in contrast with the first error function. Despite this advantage, this error function does not ensure the localisation of the experts, which is the key factor in the efficiency of ME. Therefore, Jacobs et al. introduced a new error function based on the negative log probability of generating the desired output vector, assuming a mixture of Gaussian models:

$$E_{ME} = -\log \sum_j g_j \exp\left(-\frac{1}{2}\left(y - O_j\right)^T \Sigma^{-1} \left(y - O_j\right)\right) \tag{3}$$

where Σ presents the covariance matrix in mixture model. Assuming the identity covariance matrix ($\Sigma = I$) for this model (Jacobs et al. 1991b), the above error function can be expressed as:

$$E_{ME} = -\log \sum_j g_j \exp\left(-\frac{1}{2}\left(y - O_j\right)^T \left(y - O_j\right)\right) \tag{4}$$

To evaluate this error function, its deviation with respect to *ith* expert (4) is analysed:

$$\frac{\delta E_{ME}}{\delta O_i} = -\left[\frac{g_i \exp\left(-\frac{1}{2}\left(y - O_i\right)^T \left(y - O_i\right)\right)}{\sum_j g_j \exp\left(-\frac{1}{2}\left(y - O_j\right)^T \left(y - O_j\right)\right)}\right]\left(y - O_i\right) \tag{5}$$

In this term, similar to the previous error function, the learning of each expert is based on its individual error. Moreover, the weight-updating factor for each expert is proportional to the ratio of its error value to the total error. These two features in the proposed error function that cause the localisation of each expert in their corresponding subspace eliminate the deficiencies of previous error functions. Thus, the ME method has better efficiency with this error function.

In addition, a gating network is used to complete a system of competing local experts. The gating network allows the mixing proportions of the experts to be determined by learning a partition of input space and trusts one or more expert(s) in each of these partitions. The learning rule for the gating network attempts to maximize the likelihood of the training set by assuming a Gaussian mixture model in which each expert is responsible for one component of the mixture (Dailey and Cottrell 1999).

2.3 Mixture of MLP-experts training algorithm

One of the most applied implementations for ME is mixture of MLP-experts (MME) (Waterhouse 1997; Nguyen 2006; Ebrahimpour 2007). In the version of MME, the MLP is used for the experts and gating network, instead of linear networks to improve the performance over a conventional ME. In this implementation, Each expert network is an MLP network with one hidden layer that computes an output O_i as a function of the input vector, x and weights of hidden and output layers and a sigmoid activation function. The weights of MLPs are learned using the error back-propagation, BP, algorithm, in order to maximize the log likelihood of

the training data given the parameters. For each expert i, the weights are updated according to the following rules:

$$\Delta w_y = \eta_e h_i (y - O_i)(O_i(1 - O_i))O_{h_i}^T \tag{6}$$

$$\Delta w_h = \eta_e h_i w_y^T (y - O_i)(O_i(1 - O_i))O_{h_i}(1 - O_{h_i})x_i \tag{7}$$

where η_e is learning rate for the experts. w_h and w_y are the weights of input to hidden and hidden to output layer for the experts, respectively. O_{hi}^T is the transpose of O_{hi}, the outputs of the hidden layer of expert. The first term of derivation of ME error function (Eq. 5) is shown with h_i (Eq. 8). This term can be considered as an estimation of the posterior probability that expert i can generate the desired output y:

$$h_i = \frac{g_i \exp\left(-\frac{1}{2}\|y - O_i\|^2\right)}{\sum_j g_j \exp\left(-\frac{1}{2}\left\|y - O_j\right\|^2\right)} \tag{8}$$

According to the ME error function (Eq. 4), the error function of the gating network can be written as:

$$E_G = \frac{1}{2}\|(h - Og)\|^2 \tag{9}$$

where $h = [h_i]_{i=1}^N$. Based on this error function, the weights of the gating network in the MME method are determined using the BP error algorithm according to the following rules:

$$\Delta w_{y,g} = \eta_g (h - O_g)(O_g(1 - O_g))O_{h,g}^T \tag{10}$$

$$\Delta w_{h,g} = \eta_g w_{y,g}^T (h - O_g)(O_g(1 - O_g))O_{h,g}\left(1 - O_{h,g}\right)x_i \tag{11}$$

where η_g is the learning rate, and $w_{h,g}$ and $w_{y,g}$ are the weights of the inputs to hidden and hidden to output layers of the gating network, respectively. $O_{h,g}^T$ is the transpose of $O_{h,g}$, the outputs of the hidden layer of the gating network.

In this learning procedure, the expert networks compete for each input pattern, while the gate network rewards the winner of each competition with stronger error feedback signals. Thus, over time, the gate partitions the input space in response to the expert's performance. The training step in the MME method is illustrated in Fig. 1.

For the testing step in the MME method, the gating network assigns a weight g_i to each of the experts' output, O_i. The softmax function applied on the gating network outputs leads to a higher extent of diversity. The g_i can be interpreted as estimates of the prior probability that expert i can generate the desired output y. The gating network is composed of two layers: the first layer is an MLP network, and the second layer is a softmax nonlinear operator. Thus the gating network computes, O_g which is the output of the MLP layer of the gating network, then applies the softmax function to achieve:

$$g_i = \frac{\exp\left(O_{g,i}\right)}{\sum_{j=1}^N \exp\left(O_{g,j}\right)} \qquad i = 1, \ldots, N \tag{12}$$

where N is the number of expert networks, so g_i is nonnegative and sum to 1.

Finally, to combine the experts' outputs, the gate assigns a weight g_i as function of x to each of expert's output O_i, and the final mixed output of the ensemble is:

$$O_T = \sum_{i=1}^{N} O_i g_i \tag{13}$$

Figure 2 shows the testing step in MME method.

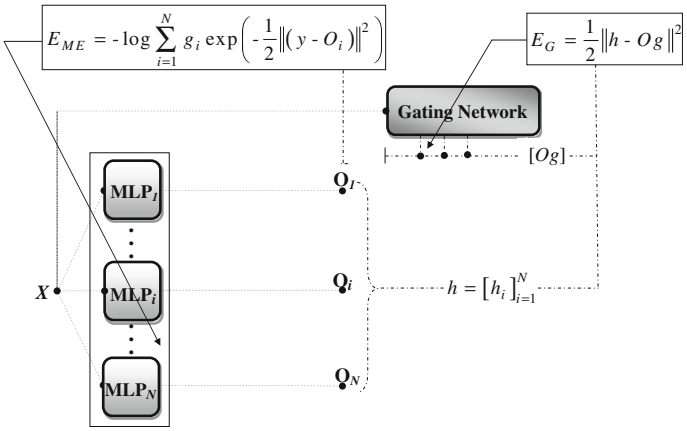

Fig. 1 Diagram for simultaneous training of the experts and gating network through the error functions of MME method is shown in Fig. 1. The experts compete to learn the training patterns, and the gating network mediates the competition

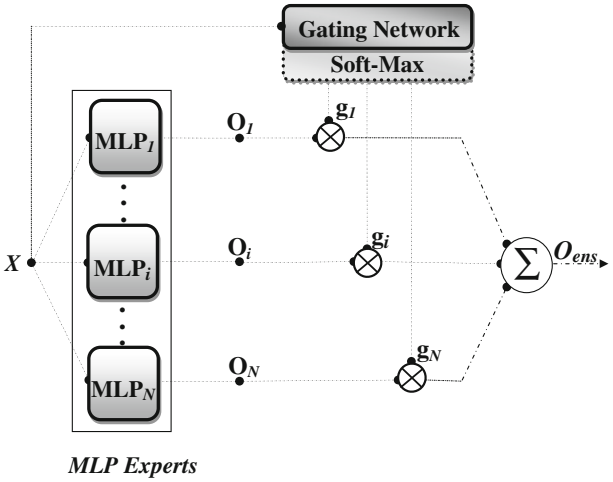

Fig. 2 Diagram for the testing step in the MME method. In this step, the input x is given to the MME experts and gating network, simultaneously and soft-max function is applied on the outputs of the gating network. The final output of ensemble system is calculated based on the weighted averaging of base MLP experts

As a result of good classification performance and transparency of the MME method, it has been widely employed in many applications (Ebrahimpour et al. 2011a,c; Ubeyli et al. 2010) since Jacobs' proposal (Jacobs et al. 1991a,b). Considering the types of learners and training algorithms employed in the learning of experts and gating, also regarding the way that gating involve in the dividing problem space, several works have been reported in both of MILE and MELE groups which are surveyed in the next section.

3 Categorisation of ME literature

Jacobs et al. (1991b) proposed an ensemble method based on the D&C principle, called ME. Approaches using D&C in the ME method were implemented in different ways, and

several methods were developed based on this model, which used different strategies to divide the problem between the experts (Jacobs et al. 1991a; Gutta et al. 2000; Tang et al. 2002; Ebrahimpour et al. 2011b). To survey and analyse these methods more clearly, we present a categorisation of the ME literature based on this difference. Various ME implementations were classified into two groups, according to the partitioning strategies used and both how and when the gating network is involved in the partitioning and combining procedures.

The conventional ME (Jacobs et al. 1991b) and the extensions of this method (Jacobs et al. 1991a; Kim et al. 2003; Ebrahimpour et al. 2008a) stochastically partition the problem space into a number of subspaces using the special employed error function (Eq. 4), and experts become specialised in each subspace. These methods use a gating network to manage this process, which are trained together with the experts. In this group, the gating network, considering the differences in the experts' efficiencies in the different sub-spaces, simultaneously co-operate in partitioning the problem when training the experts. In the second group, the problem space is explicitly partitioned by the clustering method before the experts' training process starts, and each expert is then assigned to one of these sub-spaces. Based on the implicit problem space partitioning using a tacit competitive process between the experts, we call the first group the mixture of implicitly localised experts (MILE), and the second group is called mixture of explicitly localised experts (MELE), as it uses pre-specified clusters. Below, we first review the methods in the MILE group before describing the methods in the MELE group.

3.1 Mixture of implicitly localised experts

In the conventional ME proposed by Jacobs et al. (1991b), during a competitive learning process, a number of separate experts learn to handle different but overlapped subsets of training data. Throughout learning, a gating network decides which of the experts should be used for each training case; it rewards the expert(s) with the best performance with stronger error feedback signals. Simultaneously, the gating network partitions the input space according to the experts' performance and allocates a subspace to each expert to learn.

In the MILE group, several methods attempt to improve ME task decomposition by modifying the error function of the gating network (Jacobs et al. 1991a; Hansen 1999; Ubeyli et al. 2010). Jacobs investigated the procedure of task decomposition through competition in ME method (Jacobs et al. 1991a). The authors discussed that in the problem space partitioning in ME method, there is a trade-off between selection and fusion strategies to achieve desirable task decomposition. According to analysis of bias-variance-covariance decomposition of error, it was found that selection approach reduce bias and covariance terms via localising the experts in sub-problems while fusion approach leads to variance reduction and desirable solution is acquired in optimum balance between the two terms (Kuncheva 2002; Jacobs 1997). In order to achieve a better balance in ME method, Jacobs et al. proposed an extended error function for gating network including two terms: selection term and fusion term, which in the training phase, gating network learns in the switching condition between these two error terms (Jacobs et al. 1991a). while if, on a given training pattern, the system 's performance is significantly better than it has been in the past, then the weights of gating network are adjusted to make the output corresponding to the winning expert network increase towards 1 and the outputs corresponding to the losing expert networks decrease towards 0 leading to the selection approach. Alternatively, if the system's performance has not improved, then the gating network's weights are adjusted to move all of its outputs towards some neutral value which yields fusion strategy. This extension allows the ME method to make better

switch between selection and fusion which yields better performance in comparison with conventional ME.

Hansen (1999, 2000) reported a potential problem that may occur in the competitive ME learning procedure, the zero co-efficient problem. Unfavourable initial parameters in the ME may cause this problem. In this condition, due to the low performance of the unfavourable initialized experts through the competitive, the gating network assigns near-zero weights to them, which cause elimination from the competitive learning process. To solve this problem, Hansen suggested adding a term to the error function of the gating network that leads the ensemble to fusion and averaging situations. Although this idea solves the zero-coefficient problem, it is not suitable for the ME method, as it is against the principle of localising and diversifying the experts and causes the ME method not to reach its ideal performance.

Different types of learners and training algorithms were also employed for the experts and gating network learning in the MILE (Waterhouse et al. 1996; Hansen 1999; Ebrahimpour et al. 2008d). In the conventional ME, the experts and gating network were linear classifiers; however, for more complex classification tasks, the experts and gating network could be of more complicated types (Gutta et al. 2000). For instance, Ebrahimpour et al. (2007) proposed a face detection method in which the MLPs were used in forming the expert and gating network to improve the face detection accuracy.

The methods discussed above use gradient-decent based training algorithms, but the following methods use other training algorithms for the experts and gating networks. In the training procedure, ME method attempts to achieve two goals: (1) for a given expert, find the optimal gating function, and (2) for a given gating function, train each expert to achieve maximal performance on the distribution assigned to it by the gating function. This decomposition of the learning task motivates an Expectation Maximization (EM) algorithm, though simultaneous training was also used. Jordan and Jacobs (1994) extended their method to the so-called Hierarchical mixture of experts in which each component of ME is replaced by a ME method. In Jordan and Jacobs (1994), authors indicated that the gating network performs a typical multiclass classification task (Mangiameli and West 1999; Viardot et al. 2002). EM (Chen et al. 1999) algorithm was introduced to the ME architecture in order that the learning process is separated in a way that fits well with the modular structure. Since the EM algorithm learns only the cluster centroids and not intermediary points, it will not work well on non linear examples. In Guler and Ubeyli (2005), Hong and Harris (2001), ME is used with MLPs experts for medical diagnostic systems and ECG beats classification, respectively. These papers illustrated the use of modified ME structure to guide model selection for classification of electrocardiogram beats with diverse features. EM algorithm was used for training the proposed method, so that the learning process was decoupled in a manner that fits well with the modular structure.

3.2 Mixture of explicitly localised experts

Researchers have proposed MELE methods since 2000, which generally have better performance than the MILE methods (Gutta et al. 2000; Tang et al. 2002). Although these methods have the same structure as the MILE methods, including some experts and a gating network, they work differently as the partitioning mechanism. Unlike MILE, which stochastically partitions the input space and specialises each expert network on nested and stochastic input space regions, MELE methods partition the input space into more separable spaces, and each expert is then specialised on a pre-specified subspace with altered learning rules of a conventional ME. As in the previous group, several methods are developed in the MELE group that use different learning systems for the experts and gating network.

Tang et al. attempted to explicitly localise the experts by applying a cluster-based pre-processing step to partition the input space for the experts (Tang et al. 2002). They used Self-Organising Maps as the gating network to partition the input space according to the underlying probability data distribution. As a result, they achieved a better generalisation ability with more parameter setting stability. Nevertheless, as they argued at the end of the paper, the proposed method was designed for only binary and low-dimensional problems.

Goodband et al. presented a new algorithm based on ME method to incorporate photon scatter to design compensators for intensity modulated radiation therapy (Goodband et al. 2006). The algorithm utilizes the fuzzy C-means clustering algorithm to partition data before training of the experts commences. Each MLP expert is trained on a specified overlapped subset of data. A reduction in the size of training set also allows the Levenberg–Marquardt algorithm to be implemented. ME is also controlled by a Radial-Basis Function gating network, which is designed using the centroid of each subset as a centre for each basis function.

Nguyen et al. introduced a novel method based on the principles of both Cooperative Co-evolution and mixture of experts (Nguyen et al. 2006). They showed that their method can automatically decompose problems into different regions of the input space, and assign the experts to these distinct regions. This Cooperative Co-evolution layer allows better exploration of the weight space, and hence, an ensemble with better performance was achieved.

Ebrahimpour et al. (2011a,c) proposed a view-independent face recognition system using ME by manual decomposition of the face view space into specific angles (views) (Ebrahimpour et al. 2008d). In this method, they do not rely on the unsupervised partitioning of the face space by ME, in which, similar to Self-Organizing Maps, the experts' areas of specialization are autonomously clustered. Instead, using teacher-directed learning (Kamimura 2004), the face space is divided in two conditions into a number of overlapping or disjoint subspaces and each expert being specialized on its subspace (Ebrahimpour et al. 2008b,c). Nevertheless, the proposed method is efficient in 2D face recognition and, as argued by the authors, extending this approach to other classification problems and applications could be challenging and not always possible.

3.3 Comparing MILE with MELE

This section compares MILE and MELE, discussing their advantages and disadvantages.

MILE methods divide the problem space implicitly between the experts. The implicit partitioning of the problem space has some drawbacks:

1. Problem partitioning in this method is based on different expert performance in different regions, which originates from different initial weights. This partitioning type is not efficient for the ME method, as it may leads to complex and nested partitions, and thus, the gating network cannot model it well (Tang et al. 2002).
2. In the MILE, one or more expert(s) may be eliminated from the competitive ME learning process, according to the abovementioned zero-coefficient problem (Hansen 2000).

To overcome these problems in the MILE methods, researchers have several approaches in the MELE group that attempted to partition the problem space explicitly. The MELE methods use prior knowledge to divide the task that the system would be required to perform. This approach assesses the significance of using a priori problem decomposition between the experts. This decomposition can be acquired by some criteria, e.g., clustering. Most MELE methods use a clustering step to partition the initial input space, and different partitions may

be forwarded to different experts. Tang et al. (2002) investigated such a scheme and claimed that the MELE can provide some advantages:

1. Dividing the input space into partitions according to the underlying probability data distribution can solve the problem of complex and nested partitioning in MILE.
2. The partitioning step can explicitly determines the ME architecture, as the same number of experts and clusters is selected.
3. Due to the clustering step, each expert selectively learns the corresponding region that leads to a clearer distinction between experts' responsibilities, resulting in a better generalisation ability.

In contrast, there are three unfavourable consequences of this scheme:

1. As Jacobs et al. mentioned, though domain knowledge may be useful in suggesting a priori decomposition of a task, the boundaries between subtasks are rarely explicitly marked in the data presented to the experts. Moreover, the optimal allocation of experts to subtasks depends not only on the nature of the task but also on that of the learner (Jacobs et al. 1991a). MELE did not account for this, which may be considered an advantage of MILE over the MELE methods.
2. If the interaction among NN experts in the ensemble were missing, as in pure explicit problem space decomposition, some portions of the task might remain unsolved and the bias term does not reduce significantly with training. The MILE methods overcome this problem by facilitating interaction using the competitive learning process of conventional ME.
3. The cluster-based partitioning does not consider the information regarding the data class label and may lead to unbalanced class partitioning into clusters, which is not desirable for NN experts.

In addition to categorising ME methods in these classes, as mentioned above, problem space partitioning in ME is based on dynamic switching between the selection and fusion or in other words, on the balance between hard splits of problem versus soft splits (Kuncheva 2004). To more accurately analyse this balance, Jacobs (1997) investigated the error decomposition of the ME method into bias-variance-covariance terms. As this error decomposition tends towards the selection strategy, mostly in the MELE methods, it reduces bias and covariance terms; tending towards the fusion strategy, mostly in the MELE methods, leads to variance reduction. As there is a trade-off between these constituent generalisation error terms, it is not possible to minimise them simultaneously. To reach the minimum generalisation error, a system should achieve optimal balance between these terms.

The discussed advantages and limitations of the MILE and MELE groups are summarized in Table 1.

As in the above comparison of MILE and MELE, the balance of bias-variance-covariance trade-off and the similar balance between selection and fusion also show the complementary features of these two categories.

4 Comparing ME with other popular combining methods

In this section, the popular combining methods are first shortly reviewed and compared in terms of used partitioning strategy. Next, the hybrid combining methods that employ the strengths of ME to improve their performance are investigated and some suggestions are proposed to extend these hybrid combining methods.

Table 1 Summary of the advantages and limitations of the MILE and MELE methods in creation of NN experts

Two groups of ME	Creation of NN experts	
Mixture of implicitly localised experts	Advantage	Considering not only the nature of the task but also that of the learners. Tending towards the fusion strategy that reduces the variance error term.
	Disadvantage	The complex and nested partitions for the experts that can not be modelled efficiently by the gating network. Probable elimination of the experts from the competitive ME learning process, according to the zero-coefficient problem.
Mixture of explicitly localised experts	Advantage	Considering the probability data distribution for problem partitioning can solve the problem of complex and nested partitioning in MILE. Determination of the ME architecture, as the same number of experts and clusters. A clearer distinction between experts' responsibilities, due to the clustering step. Tending towards the selection strategy that reduces bias and covariance error terms.
	Disadvantage	The optimal allocation of experts to subtasks depends not only on the nature of the task but also on that of the learners, not to be considered in MELE. Missing the interaction among NN experts in pure explicit problem space decomposition and probably remaining some portions of the task unsolved. Not considering the information regarding the data class label that may lead to unbalanced class partitioning into clusters.

4.1 Combining methods, in terms of partitioning strategy

Both theoretical and experimental studies (Tumer and Ghosh 1996) have shown that more improved generalization ability can be obtained by combining the outputs of NN experts which are accurate and their errors are negatively correlated (Jacobs 1997; Hansen 2000). It was shown that in contrast to systems that produce unbiased experts whose estimation errors are uncorrelated, ME architectures produce biased experts whose estimates are negatively correlated (Jacobs 1997).

As mentioned before, Sharkey et al. argued that training NNs using different training sets is likely to produce more uncorrelated errors than other approaches (Sharkey and Sharkey 1997). Common combining methods use various strategies to produce different training sets for training individual NNs. In bagging (bootstrap aggregating), the training set is randomly sampled k times with replacement, producing k training sets with sizes equal to the original training set for k different experts. Boosting, on the other hand, trains the ensemble of NNs sequentially by adaptively changing the distribution of the training set based on the accuracy of the previously created NNs. Similar to bagging, the NNs are generated by resampling the

training set; while in the resampling mechanism of boosting, the training samples that were wrongly predicted by former NNs will play more important roles in the training of later NNs.

While bagging and boosting create explicitly different training sets for different NNs by probabilistically changing the distribution of the original training data, NCL and ME implicitly create different training sets by encouraging different NN experts to learn different parts or aspects of the training data (Liu and Yao 1999b). The key idea behind the NCL is to introduce a correlation penalty term to the error function of each individual NNs so that each NN minimizes its error together with the correlation of the ensemble. The introduced error function in the NCL encourages the individual NNs in an ensemble to learn different parts or aspects of a training data simultaneously and interactively, so that the ensemble can learn the whole training data better.

After reviewing common combining methods and their approaches of partitioning training set between NN experts, we present the motivations for integrating them in a hybrid combining method. As the explicit and implicit approaches of partitioning training set have complementary strengths and limitations, previous researches have attempted to combine their features in hybrid approaches (Waterhouse 1997; Islam et al. 2008; Ebrahimpour et al. 2011a,c). In the next section, we review the hybrid combining methods in which the complementary features of ME are incorporated in the other combining methods to address their limitations.

4.2 Hybrid combining methods, by using the complementary features of ME

In this section, the hybrid approaches that attempted to incorporate the complementary features of in the other combining methods are first reviewed. Next, we compare the ME and NCL method, discussing their advantages and limitations against each other and presenting a motivation for combining the both methods in a hybrid combining method.

Waterhouse and Cook (1997) and Waterhouse (1997) attempted to combine the features of boosting and ME. They proposed two approaches to address the limitations of each method and overcome them by combining elements of the other method. The first approach may be viewed as an improved ME that initializes the partitioning of the data set for assignment to different experts in a boost-like manner. Because boosting encourages classifiers to become experts on patterns that previous experts disagree on, it can be successfully used to split the data set into regions for each expert in the ME method, thus ensuring their localisation. The second approach may be viewed as an improved variant of the boosting algorithm, in which the main advantage is the use of a dynamic combination method for the outputs of the boosted networks.

Avnimelech and Intrator (1999) extended Waterhouse's work and proposed a new dynamically boosted ME method. They analysed the learning mechanism of two ensemble algorithms: boosting and ME. The authors discussed the advantages and weaknesses of each algorithm and reviewed several ways in which the principles of these algorithms can be combined to achieve improved performance. Furthermore, they suggested a flexible procedure for constructing a dynamic ensemble based on the principles of these two algorithms. The proposed ensemble method employs a confidence measure as the gating function, which determines the contribution of each expert to the ensemble output. This proposed method outperformed both static approaches previously proposed by Waterhouse and Cook (1997). The advantages of the proposed method are that it uses a flexible procedure for constructing ensemble in an incremental structure and also a flexible gating function. However, the authors did not present and discuss the used and also the available flexible parameters and their effects on the performance of the proposed method. The other disadvantage is that the suggested algorithm for constructing ensemble in an incremental structure has a heavy computational

load because of re-partitioning and so re-training of whole ensemble members after adding a new expert.

As a result of using boosting method for initialization of ME structure in the previous approach (Waterhouse and Cook 1997), the problem space is not distributed well-balancedly between the experts and thus, the initialization procedure does not meet its goal of training the experts in different parts of training set. In order to solve this problem, Ebrahimpour et al. (2012) proposed an approach in which boost-wise partitioning procedure was modified to train the base networks in ensemble on different balanced subsets of training set. In this paper, a pre-loading procedure was suggested included two separate steps: confidence-based boost-wise partitioning step and initialization step. In the first step, both of the error and confidence measures are used as the difficulty criteria in the boost-wise partitioning of problem space. Using confidence criteria in addition to the error measure provides more flexibility in filtering procedure so that this partitioning step can be conducted in the way that yields to balance partitioning of problem space. According to the nature of implementation, the proposed method was called Boost-wise Pre-loaded ME.

After reviewing previous researches attempted to combine the features of boosting and ME methods, we investigate the NCL and ME methods in comparison with each other to provide a novel idea to present an improved version of ME using the feature of NCL.

As mentioned before, combining systems have two major components. Regarding the first component, the creation of individual NN experts, NCL has almost the better efficiency. Its superiority comes from its use of a regularization term that provides a convenient way to balance the bias-variance-covariance trade-off and thus improves the generalization ability, whereas ME does not include such control over the trade-off (Liu and Yao 1999a). However, from the other viewpoint, ME has an partly advantage over NCL in the creation of individual NN experts. This advantage comes from the error function of ME in which the learning of each NN expert is based on its individual error, while in the NCL all individual NNs are concerned with the whole ensemble error (Islam et al. 2003).

In contrast, ME provides a better approach for the second component of combining systems, the combination of base NN experts. One of the advantages of ME over other combining methods is its distinct technique for combining the outputs of the base experts. ME uses a trainable combiner that, according to the input x, dynamically selects the best expert(s) and combines their outputs to create the final output (Kuncheva 2004). The combining function of ME includes a dynamic weighted average in which the local competence of the experts with respect to the input are estimated by the weights produced by the gating network. The outputs of all experts responsible for input x are then combined. But in NCL method, the static combining methods such as: average or winner take all method were used to combine the NCL experts (Liu and Yao 1999a), while these static methods does not have the capability to model the local competence of NCL experts.

The mentioned advantages and limitations of the ME and NCL methods in each component of a combining system are summarized in Table 2.

As it is clear from the analysis of the features of both methods and their advantages and disadvantages, the two methods have complementary features.

Characterization of both methods showed that they have different but complementary features. Based on the similar ensemble structures and strategies used in both the NCL and ME methods and due to their complementary features, several researches attempted to combine the principles of both should address their limitations and overcome them by combining elements of the other method. Ebrahimpour et al. (2011a) proposed an improved version of NCL method in which the capability of gating network, as the combining part of ME, is used to combine the base NNs in the NCL ensemble method. The gating network

Table 2 Summary of the advantages and limitations of the ME and NCL methods in each component of a combining system

Ensemble method	Combining component	Part 1 Creation of NN experts	Part 2 Combination of NN experts
Mixture of experts	Advantage	• Each NN expert is trained based on its individual error and localised in their corresponding subspace based on D&C approach. • It can produce individual NNs whose errors tend to be probably negatively correlated.	• A gating network is used to compute the weights of combining dynamically from the inputs, according to the local expertise of each expert.
	Disadvantage	• There is no control over the bias-var-cov trade-off.	–
Negative correlation learning	Advantage	• There is a regularization term that provides a convenient way to balance the bias-var-cov trade-off, thus improving the generalization ability. • It can produce individual NNs whose errors tend to be near optimum negatively correlated.	–
	Disadvantage	• All individual NNs are concerned with the whole ensemble error.	• Previously used static combiner methods does not have the capability to model the local competence of NCL experts.

provides a way to support this needed functionality for combining the NCL experts. So the proposed method was called Gated-NCL. As the other viewpoint, Masoudnia et al. (2012) presented an approach to introduce the advantage of NCL into the training algorithm of ME, i.e., Mixture of Negatively Correlated Experts. In this proposed method, the capability of a control parameter for NCL is incorporated in the error function of ME, which enables its training algorithm to establish better balance in bias-variance-covariance trade-off and thus improves the generalization ability.

5 Conclusion

ME, as one of the most interesting combining methods, was reviewed and investigated in this paper. ME is established based on D&C principle and a gating network supervise problem space partitioning between the experts. In earlier works on ME, different strategies were developed to divide the problem space between the experts. After reviewing ME's place in the combining taxonomies, to survey and analyse the ME-based methods more clearly, we presented our proposed categorisation of the ME literature that classifies these methods into two categories based on the differences in the partitioning used: Mixture of

Implicitly Localised Experts (MILE) and Mixture of Explicitly Localised Experts (MELE). In the MILE methods, the problem space is stochastically partitioned into a number of sub-spaces, experts become specialized on each subspace. The gating network, during the training of the experts, with respect to difference in the experts' efficiencies in the different sub-spaces co-operate and manage the partitioning of problem simultaneously. In the MELE methods, the problem space is explicitly partitioned by the clustering method before the experts' training process starts and, each expert is then assigned to one of these pre-specified sub-spaces. The properties of two groups were analysed in comparison with each others. As discussed comparison of MILE and MELE, the balance of bias-variance-covariance trade-off and the similar balance between selection and fusion, also show the complementary features of these two categories. Moreover, ME was compared with two other popular combining methods, including boosting and NCL. Characterization of ME and the other methods showed that they have complementary strengths and limitations. Considering the complementary properties of ME compared to other combining methods and, moreover, of the MILE versus MELE, the question arises as to whether their integration can lead to the improved and more powerful combining schemes. We studied this question, and based on the complementary features of these methods, some suggestions were proposed for future research directions in the next section.

6 Future works

Analysing the different advantages and disadvantages of the MILE and MELE methods and moreover, of the ME compared to other combining methods, some intuitions are proposed as future research directions to achieve a better trade-off between the bias, variance and covariance terms of generalisation error in the ensemble or, similarly, better balance between the selection and fusion strategies. To meet these objectives, some approaches can be applied in MILE or MELE, including the following ideas:

1. Incorporating a free parameter in the MILE or MILE training algorithms that can control this trade-off, similar to the lambda factor in the NCL method or the temperature factor suggested by Jacobs (1997), can enhance the generalisation performance.
2. If the MELE training algorithm is modified to encourage the experts to have more cooperation, especially in the overlapping regions between the expertise areas of two or more experts, this would lead to a greater generalisation ability.
3. The ME structure is well suited for incorporating prior knowledge to bias the nature of the decomposition to be performed. This property could be used to combine the complementary features of MILE and MELE groups. If a two-step hybrid system could be designed in which, first a priori data acquired by clustering was used to initialise the experts on different distributions and in the second step, as in MILE methods, the conventional ME learning algorithm is performed. This hybrid approach may avoid the potential MILE problems. Due to the initializing step, the disadvantages of the nested and complex problem space partitioning caused by initial random weights may be decreased, and network initialisation in the first step makes each expert network also can learn considerable problem sub-spaces, which avoids the zero-coefficient problem. Moreover, an interesting future direction is using other methods rather than clustering to partition the training set based on different parameters, including confidence or other difficulty criteria.

In addition to above suggestions, regarding the previously developed methods based on the ME model in which BP training algorithm were used for convergence, the identity constraint

was applied to the covariance matrix in the mixture models. To overcome this limitation and improve the efficiency for the ME-based methods, a dynamic process can be employed to optimize the covariance matrix of mixture models during the ME training algorithm. The global heuristic optimization approaches such as evolutionary or swarm-based optimization algorithms are good candidates for this application.

References

Alpaydin E, Jordan MI (1996) Local linear perceptrons for classification. IEEE Trans Neural Netw 7(3): 788–792

Avnimelech R, Intrator N (1999) Boosted mixture of experts: an ensemble learning scheme. Neural Comput 11(2):483–497

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Chen K, Xu L, Chi H (1999) Improved learning algorithms for mixture of experts in multiclass classification. Neural Netw 12(9):1229–1252

Dailey MN, Cottrell GW (1999) Organization of face and object recognition in modular neural network models. Neural Netw 12(7–8):1053–1074

Dietterich T (2000) Ensemble methods in machine learning. Multiple classifier systems, Cagliari, Italy. Springer, LNCS, pp 1–15

Ebrahimpour R (2007) View-independent face recognition with mixture of experts. PhD thesis, Institute for studies in Theoretical Physics and Mathematics (IPM)

Ebrahimpour R, Kabir E, Yousefi MR (2007) Face detection using mixture of MLP experts. Neural Process Lett 26(1): 69–82. doi:10.1007/s11063-007-9043-z

Ebrahimpour R, Kabir E, Esteky H, Yousefi MR (2008a) A mixture of multilayer perceptron experts network for modeling face/nonface recognition in cortical face processing regions. Intell Autom Soft Comput 14(2):151–162

Ebrahimpour R, Kabir E, Yousefi MR (2008b) Teacher-directed learning in view-independent face recognition with mixture of experts using overlapping eigenspaces. Comput Vis Image Underst 111(2): 195–206. doi:10.1016/j.cviu.2007.10.003

Ebrahimpour R, Kabir E, Yousefi MR (2008c) Teacher-directed learning in view-independent face recognition with mixture of experts using single-view eigenspaces. J Franklin Inst 345(2): 87–101. doi:10.1016/j.jfranklin.2007.06.004

Ebrahimpour R, Kabire E, Esteky H, Yousefi MR (2008d) View-independent face recognition with mixture of experts. Neurocomputing 71(4–6): 1103–1107. doi:10.1016/j.neucom.2007.08.021

Ebrahimpour R, Nikoo H, Masoudnia S, Yousefi MR, Ghaemi MS (2010) Mixture of MLP-experts for trend forecasting of time series: a case study of the Tehran stock exchange. Int J Forecast

Ebrahimpour R, Arani SAAA, Masoudnia S (2011a) Improving combination method of NCL experts using gating network. Neural Comput Appl 1–7. doi: 10.1007/s00521-011-0746-8.

Ebrahimpour R, Kabir E, Yousefi MR (2011b) Improving mixture of experts for view-independent face recognition using teacher-directed learning. Mach Vis Appl 22(2):421–432

Ebrahimpour R, Nikoo H, Masoudnia S, Yousefi MR, Ghaemi MS (2011c) Mixture of MLP-experts for trend forecasting of time series: a case study of the Tehran stock exchange. Int J Forecast 27(3):804–816. doi:10.1016/j.ijforecast.2010.02.015

Ebrahimpour R, Sadeghnejad N, Arani SAAA, Mohammadi N (2012) Boost-wise pre-loaded mixture of experts for classification tasks. Neural Comput Appl:1-13. doi:10.1007/s00521-012-0909-2

Goodband JH, Haas OCL, Mills JA (2006) A mixture of experts committee machine to design compensators for intensity modulated radiation therapy. Pattern Recogn 39(9): 1704–1714. doi:10.1016/j.patcog.2006.03.018

Guler I, Ubeyli ED (2005) A mixture of experts network structure for modelling Doppler ultrasound blood flow signals. Comput Biol Med 35(7): 565–582. doi:10.1016/j.compbiomed.2004.04.001

Gutta S, Huang JRJ, Jonathon P, Wechsler H (2000) Mixture of experts for classification of gender, ethnic origin, and pose of human faces. IEEE Trans Neural Netw 11(4):948–960

Hansen JV (1999) Combining predictors: comparison of five meta machine learning methods. Inform Sci 119(1–2):91–105

Hansen JV (2000) Combining predictors: meta machine learning methods and bias/variance and ambiguity decompositions. Computer Science Dept., Aarhus Univ, Aarhus

Hong X, Harris CJ (2001) A mixture of experts network structure construction algorithm for modelling and control. Appl Intell 16(1):59–69
Islam MM, Yao X, Murase K (2003) A constructive algorithm for training cooperative neural network ensembles. IEEE Trans Neural Netw 14(4):820–834
Islam MM, Yao X, Nirjon SMS, Islam MA, Murase K (2008) Bagging and boosting negatively correlated neural networks. IEEE Trans Syst Man Cybern B 38(3): 771–784. doi:10.1109/Tsmcb.2008.922055
Jacobs RA (1997) Bias/variance analyses of mixtures-of-experts architectures. Neural Comput 9(2):369–383
Jacobs RA, Jordan MI, Barto AG (1991) Task decomposition through competition in a modular connectionist architecture—the what and where vision tasks. Cogn Sci 15(2):219–250
Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Comput 3(1):79–87
Jordan MI, Jacobs RA (1994) Hierarchical mixtures of experts and the Em algorithm. Neural Comput 6(2):181–214
Kamimura R (2004) Teacher-directed learning with Gaussian and sigmoid activation functions. In: Springer, New York, pp 530–536
Kecman V (2001) Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. The MIT press, Cambridge
Kim SP, Sanchez JC, Erdogmus D, Rao YN, Wessberg J, Principe JC, Nicolelis M (2003) Divide-and-conquer approach for brain machine interfaces: nonlinear mixture of competitive linear models. Neural Netw 16(5–6): 865–871. doi:10.1016/S0893-6080(03)00108-4
Kittler J, Hatef M, Duin RPW, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal 20(3): 226–239
Kotsiantis S (2011a) Combining bagging, boosting, rotation forest and random subspace methods. Artif Intell Rev 35(3):1–18. doi:10.1007/s10462-010-9192-8
Kotsiantis S (2011b) An incremental ensemble of classifiers. Artif Intell Rev 36(4):1–18. doi:10.1007/s10462-011-9211-4
Kotsiantis S, Zaharakis I, Pintelas P (2006) Machine learning: a review of classification and combining techniques. Artif Intell Rev 26(3):159–190
Kuncheva LI (2002) Switching between selection and fusion in combining classifiers: an experiment. IEEE Trans Syst Man Cybern B 32(2): 146–156. doi:PiiS1083-4419(02)00697-0
Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley-Interscience, New York
Liu Y, Yao X (1999) Ensemble learning via negative correlation. Neural Netw 12(10):1399–1404
Liu Y, Yao X (1999) Simultaneous training of negatively correlated neural networks in an ensemble. IEEE Trans Syst Man Cybern B 29(6):716–725
Lorena AC, de Carvalho AC, Gama JMP (2008) A review on the combination of binary classifiers in multiclass problems. Artif Intell Rev 30(1):19–37
Masoudnia S, Ebrahimpour R, Arani SAAA (2012) Incorporation of a regularization term to control negative correlation in mixture of experts. Neural Process Lett:1-17
Mangiameli P, West D (1999) An improved neural classification network for the two-group problem. Comput Oper Res 26(5):443–460
Nguyen MH (2006) Cooperative coevolutionary mixture of experts: a neuro ensemble approach for automatic decomposition of classification problems. University of New South Wales, New South Wales
Nguyen MH, Abbass HA, Mckay RI (2006) A novel mixture of experts model based on cooperative coevolution. Neurocomputing 70(1–3): 155–163. doi:10.1016/j.neucom.2006.04.009
Polikar R (2006) Ensemble based systems in decision making. IEEE Circ Syst Mag 6(3):21–45
Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33(1–2): 1–39. doi: 10.1007/s10462-009-9124-7
Schapire RE (1990) The strength of weak learnability. Mach Learn 5(2):197–227
Sharkey AJC, Sharkey NE (1997) Combining diverse neural nets. Knowl Eng Rev 12(03):231–247
Tang B, Heywood MI, Shepherd M (2002) Input partitioning to mixture of experts. In: IEEE, pp 227–232
Tran TP, Nguyen TTS, Tsai P, Kong X (2011) BSPNN: boosted subspace probabilistic neural network for email security. Artif Intell Rev 35(4):1–14
Tumer K, Ghosh J (1996) Error correlation and error reduction in ensemble classifiers. Connect Sci 8(3): 385–404
Ubeyli ED (2009) Modified mixture of experts employing eigenvector methods and Lyapunov exponents for analysis of electroencephalogram signals. Expert Syst 26(4): 339–354. doi:10.1111/j.1468-0394.2009.00490.x
Ubeyli ED, Ilbay K, Ilbay G, Sahin D, Akansel G (2010) Differentiation of two subtypes of adult hydrocephalus by mixture of experts. J Med Syst 34(3): 281–290. doi:10.1007/s10916-008-9239-4

Viardot G, Lengelle R, Richard C (2002) Mixture of experts for automated detection of phasic arousals in sleep signals. In: IEEE International Conference on Systems, Man and Cybernetics, pp 551–555
Waterhouse S, Cook G (1997) Ensemble methods for phoneme classification. Adv Neural Inf Processing Syst 9:800–806
Waterhouse S, MacKay D, Robinson T (1996) Bayesian methods for mixtures of experts. Citeseer
Waterhouse SR (1997) Classification and regression using mixtures of experts. Unpublished doctoral dissertation, Cambridge University
Woods K, Kegelmeyer WP, Bowyer K (1997) Combination of multiple classifiers using local accuracy estimates. IEEE Trans Pattern Anal 19(4):405–410
Xing HJ, Hua BG (2008) An adaptive fuzzy c-means clustering-based mixtures of experts model for unlabeled data classification. Neurocomputing 71(4–6): 1008–1021. doi:10.1016/j.neucom.2007.02.010